A JOURNAL BY SISSA MEDIALAB

JCom

SPECIAL ISSUE
Science Communication in the Age of Artificial Intelligence

ARTICLE

"ChatGPT, is the influenza vaccination useful?" Comparing perceived argument strength and correctness of pro-vaccination arguments from AI and medical experts

Selina A. Beckmann, Elena Link and Marko Bachl

Abstract

Realizing the ascribed potential of generative AI for health information seeking depends on recipients' perceptions of quality. In an online survey (N = 294), we aimed to investigate how German individuals evaluate AI-generated information compared to expert-generated content on the influenza vaccination. A follow-up experiment (N = 1,029) examined the impact of authorship disclosure on perceived argument quality and underlying mechanisms. The findings indicated that expert arguments were rated higher than AI-generated arguments, particularly when authorship was revealed. Trust in science and the Standing Committee on Vaccination accentuated these differences, while trust in AI and innovativeness did not moderate this effect.

Keywords

AI tools in science communication; Health communication

Received: 23rd October 2024
Accepted: 28th February 2025
Published: 14th April 2025

1 ▪ Introduction

Generative artificial intelligence (AI), such as ChatGPT, has gained significant attention for its ability to educate users [Palmer & Spirling, 2023], explain complex topics in accessible language [Deiana et al., 2023; Meng et al., 2024], and allow users to ask follow-up questions [Lee et al., 2023] until they feel sufficiently informed. These functions are particularly beneficial for health information, which is often complex, uncertain, and difficult to understand [Carcioppolo et al., 2016]. AI can lower access barriers and improve comprehension [Reis et al., 2024]. Despite these benefits, concerns exist about the factual accuracy and reliability of AI-generated content, as AI chatbots are probabilistic language models rather than verified knowledge sources [Yildirim & Paul, 2024]. Nevertheless, studies indicate that AI-generated health information meets quality requirements [Ayers et al., 2023; Deiana et al., 2023; Hershenhouse et al., 2024; Song et al., 2023; Zhou et al., 2023].

Aside from expert discourses and quality assessments of AI-generated health content, public perception of quality has seldom been examined [Reis et al., 2024]. However, the users' perspective is crucial to realizing the ascribed potentials of AI health information, as the characteristics and utility of generated health content influence information-seeking behaviors [Johnson & Meischke, 1993]. Preliminary findings suggest that AI-generated information can be perceived as superior to human-produced texts [e.g., Karinshak et al., 2023; Lim & Schmälzle, 2023; Palmer & Spirling, 2023], potentially due to factors such as readability and sentiment [Lim & Schmälzle, 2023]. The superior perception of AI-generated information depends on recipients being unaware of its AI origin, which has been observed across various contexts [Ayers et al., 2023; Jakesch et al., 2019; Karinshak et al., 2023; Lim & Schmälzle, 2023; Palmer & Spirling, 2023; Reis et al., 2024]. When AI authorship is disclosed, the state of research is heterogeneous. While studies in the context of robot journalism from China and South Korea found higher quality ratings for AI-labeled content [Jung et al., 2017; Zheng et al., 2018], studies from the United States and the Netherlands [Graefe et al., 2018; Waddell, 2018; Zheng et al., 2018] and studies examining content produced by generative AI more broadly reached the opposite conclusion [e.g., Karinshak et al., 2023; Lim & Schmälzle, 2024; Palmer & Spirling, 2023; Reis et al., 2024; Teigen et al., 2024]. Previous research has proposed initial explanations for the shift in quality perception in the case of source disclosure, drawing on dual process models of information processing [e.g., Eagly & Chaiken, 1993]. Their general assumption is that heuristics and peripheral cues play a crucial role in shaping perceptions of information. In this context, cues referring to the source, such as its ascribed credibility, pre-existing general attitudes toward AI [Jakesch et al., 2019; Lim & Schmälzle, 2024], or a general aversion to algorithmic decision-making [Reis et al., 2024], serve as mental shortcuts that influence the evaluation of information quality [Ismagilova et al., 2020].

Against this backdrop, we investigate how individuals in Germany assess AI-generated health information compared to expert-provided content on the influenza vaccination. The expert-provided content refers to "medical expertise", which covers content provided by scientific institutions and organizations such as the Robert Koch Institute, rather than individual healthcare providers. To address this objective, we conducted a two-study design. First, an online survey examined how individuals assess pro-vaccination arguments generated by ChatGPT and medical experts without disclosing the authorship. Second, in an online experiment, we analyzed the influence of disclosing authorship on recipients' perception of an argument's quality. To gain further insights into the underlying mechanisms

of the effect of labeling, we considered trust in AI, science, and the Standing Committee on Vaccination (STIKO), as well as innovativeness — a personality trait reflecting individuals' propensity to adopt new technologies [Goldsmith, 2011; Goldsmith & Hofacker, 1991] — of potential users as moderators. Both studies focused on information about influenza vaccination, as vaccine hesitancy is one of the top ten threats to global health [World Health Organization, 2019], and influenza vaccines are recommended for a large proportion of the population [Robert Koch Institute, 2023].

2 · Study 1: online survey on recipients' quality assessment

The first online survey investigated how recipients perceive the quality of AI- versus expert-generated arguments on influenza vaccination when authorship is undisclosed. We conceptualized quality based on Trepte et al. [2005], who define it as a multidimensional construct including intrinsic content characteristics (e.g., accuracy, correctness, completeness) and recipients' perceptions (e.g., comprehensibility, relevance, usefulness). We focus on the subdimension of recipients' perceptions of quality. Within this dimension, we distinguish between argument strength and correctness. In line with Zhao et al. [2011], argument strength encompasses the aspects of comprehensibility, relevance, and usefulness of content. To comprehensively capture content quality, we additionally consider perceived correctness, which describes whether the content is rated as accurate [Trepte et al., 2005]. Since, to our knowledge, no prior studies have investigated perceptions of AI-generated health content versus expert-authored content in a German-language context, we derived the following research question:

> **RQ1:** How do individuals assess the strength and correctness of AI-generated versus expert-generated health arguments?

Previous research suggests that AI-generated arguments are rated more favorably than expert-authored content when authorship is not disclosed [Jakesch et al., 2019; Karinshak et al., 2023; Lim & Schmälzle, 2023; Palmer & Spirling, 2023]. Although detailed explanations for these differences are scarce, factors such as readability, linguistic style, and sentiment may contribute to these perceptions [Lim & Schmälzle, 2023]. AI-generated content might be perceived as clearer and easier to understand, which influences judgments of quality. Building on this state of research, we postulate that recipients ascribe a higher quality to AI-generated than expert-authored arguments when the source is not disclosed. We extend the current state of research by distinguishing between argument strength (H1) and correctness (H2). The corresponding hypotheses are:

> **H1:** The argument strength is rated stronger for arguments generated by AI than for arguments authored by medical experts when authorship is not disclosed.
> **H2:** The correctness is rated higher for arguments generated by AI than for arguments authored by medical experts when authorship is not disclosed.

2.1 ▪ *Methods*

2.1.1 ▪ *Participants and procedure*

To answer our research question and test the hypotheses, we conducted an online survey of German residents (N = 294), recruited via a regional online access panel with a heterogeneous composition. The panel members were initially recruited in 2020 by inviting a representative sample of citizens from a large southwestern city in Germany to participate in a survey via postal mail [Brettschneider & Bachl, 2020]. While participants who self-selected into the panel remained for several years and responded to our invitation, they were no longer a representative sample of the original population; however, they were still more typical than convenience samples of students or snowball samples recruited online. The participants were, on average, 56.7 years old (SD = 14.8), and 48.6% identified as female. The sample was predominantly highly educated, with 79.6% having at least a high school diploma, 15.6% having a secondary school diploma, and 3.7% having a lower secondary school diploma or no high school diploma. According to German legal regulations, this type of data collection was considered exempt from the need for ethical approval. Nevertheless, to meet ethical considerations, participants were informed about the voluntary nature of participation, their right to withdraw, anonymity, and the type of data collection at the beginning of the survey. They were also asked to give their informed consent before answering the survey. After participation, they were provided with information about the research interest and were asked to renew their informed consent.

Participants were shown eight informational texts on the advantages of influenza vaccination, four of which were generated with ChatGPT using the GPT-3 version that was publicly available at the time of data collection. We used a standard account with no specific user settings. To simulate typical usage [Karinshak et al., 2023], we engaged ChatGPT in four separate sessions with German prompts such as, “Why should I get vaccinated against influenza?” ChatGPT provided comprehensive lists of arguments, which were subsequently ranked and condensed into the five most important ones within each session. Across sessions, four recurring topics emerged as most salient: protection against serious illness, protection of the community, protection of the healthcare system, and reduction of individual risk. The research team independently assigned the AI-generated arguments from these sessions to the identified topics, ensuring intersubjective validity. For each topic, a list of arguments was created, ranging from 7 (e.g., “reduction of individual risk”) to 25 (e.g., “protection against severe illness”). A random sampling procedure was used to select one argument per topic for inclusion.

To enable content comparability, expert-authored arguments were sourced from publicly available online materials on the influenza vaccination provided by leading health institutions in Germany: the Robert Koch Institute, STIKO, and the Federal Center for Health Education. Arguments were categorized by the same topics, and one argument per topic was randomly selected. A comparison of the different arguments showed that the AI arguments tended to be shorter, had a less complicated sentence structure, and were less descriptive. All eight arguments were presented in randomized order without indicating the source. Each participant assessed all arguments, allowing for within-subject comparisons.

2.1.2 ▪ *Measures*

Argument strength. To measure how participants perceive argument strength after exposure to each argument, we used the perceived argument strength scale by Zhao et al. [2011]. The nine-item measure (e.g., "The statement is a reason for getting the influenza vaccination that is convincing. /… that is important to me.") collected responses on a 5-point Likert-type scale (see appendix A, Table 3). The internal consistency of the scale was satisfactory across the arguments (Cronbach's α = .92–.95) and allowed for the calculation of mean indices (see Table 1).

Correctness. The perceived correctness was measured in line with Kohring and Matthes [2004] using a scale of four items (e.g., "The text presents the facts as they are." or "The information given is true.") that were answered on a 5-point Likert-type scale. Based on satisfactory internal consistency (Cronbach's α = .92–.95), we calculated mean indices per argument (see Table 1).

2.1.3 ▪ *Data analysis*

We performed descriptive analysis (RQ1) and two Repeated Measures ANOVAs (RM-ANOVAs) (H1 and H2) using SPSS version 27. Given that each participant evaluated multiple arguments, RM-ANOVAs were appropriate to control for inter-individual variability. The dependent variables were the assessments of argument strength and correctness per argument. The independent variables were the different sources of arguments.

2.2 ▪ *Results*

Regarding RQ1, which examined the argument strength and correctness of the arguments provided, the descriptive results showed that both expert- and AI-generated arguments were rated as rather convincing and correct by the recipients (see Table 1). A detailed look at the individual arguments showed that the argument perceived as strongest and most correct was the medical experts' argument about the reduction of individual risks through vaccination, whereas the same AI argument received the lowest rating on argument strength and correctness. The differences were statistically significant. Both arguments about protection against serious illness received similar ratings of argument strength, but the AI argument received higher ratings of correctness than the expert argument. A similar pattern with comparable perceptions of argument strength, but better ratings for correctness for AI than for experts was observed for the arguments about the protection of the healthcare system. However, these differences were not statistically significant. The argument about the protection of the community by experts was perceived as significantly stronger and more correct than the AI version.

Regarding H1 and H2, which stated that argument strength and correctness are assessed higher for the AI arguments than for the experts' arguments, the results of the RM-ANOVAs showed that the perception of argument strength ($F(1,250) = 65.65$, $p < .001$, $\eta_p^2 = .208$, f = .26) and correctness ($F(1,268) = 25.35$, $p < .001$, $\eta_p^2 = .086$, f = .09) both significantly differed by the source of the argument. Perceived argument strength and correctness of arguments originating from medical experts were significantly higher than those of AI arguments (see Table 1). Thus, H1 and H2 were not supported.

Table 1. Descriptive results of argument strength and correctness.

	Perceived Argument Strength		**Perceived Correctness**	
Argument theme	Argument by Experts *M* (*SD*)	AI *M* (*SD*)	Argument by Experts *M* (*SD*)	AI *M* (*SD*)
Protection against serious illness	3.68 (0.90)	3.69 (0.93)	3.86 (1.00)	3.98 (1.00)
Protection of the community	3.55[a] (1.01)	3.24[a] (0.99)	3.73[d] (1.13)	3.55[d] (1.13)
Protection of the healthcare system	3.39 (1.03)	3.38 (1.02)	3.66 (1.11)	3.74 (1.06)
Reduction of individual risks	3.70[b] (0.92)	3.11[b] (1.07)	3.99[e] (1.00)	3.44[e] (1.17)
Overall	3.59[c] (0.80)	3.35[c] (0.84)	3.81[f] (0.91)	3.68[f] (0.94)

N = 294, Results of two RM-ANOVAs**;** superscript letters indicate significant differences, values sharing the same letter differ significantly; Perceived argument strength: $F(1,250) = 65.65$, $p < .001$, $\eta_p^2 = .208$, f = .26; Perceived correctness: $F(1,268) = 25.35$, $p < .001$, $\eta_p^2 = .086$, f = .09; The requirements for conducting the RM-ANOVAs were checked and found to be assessed as fulfilled.

2.3 ▪ *Discussion of the results of study 1*

Our first study examined how recipients perceive AI-generated pro-vaccination arguments compared to those authored by medical experts when authorship is undisclosed. Contrary to previous research, we could not support the notion that AI-generated information receives higher quality ratings than human-generated information [Jakesch et al., 2019; Karinshak et al., 2023; Lim & Schmälzle, 2023; Palmer & Spirling, 2023]. While both AI and medical experts provided strong and correct information from the participants' perspective, the analysis showed that overall, the experts' information received significantly more consistent and better ratings, while AI-generated arguments exhibited high variability in terms of perceived strength and correctness. This variability might explain why, despite some AI arguments being rated as highly as expert arguments, expert content is favored overall. It should be noted, however, that the effect size for correctness was rather small.

A potential methodological explanation for our contrasting findings lies in simulating everyday use by prompting the default version of ChatGPT without pre-training. Previous studies [e.g., Karinshak et al., 2023] pre-trained AI models with expert-crafted arguments, potentially enhancing the perceived quality of AI-generated content or making it more similar to human-written arguments.

3 ▪ Study 2: effects of disclosing the authorship

The second study examined how source labeling affects the assessment of argument strength and correctness, and which moderators influence this effect. Two commonly discussed moderators were considered: trust [Jung et al., 2017; Karinshak et al., 2023] and innovativeness [Jang et al., 2023; Jung et al., 2017; Khan et al., 2019].

Based on dual process models such as the heuristic-systematic model (HSM) [Eagly & Chaiken, 1993], we assume that heuristics, understood as mental shortcuts, are crucial for

attitude formation to manage one's limited cognitive capacity and enable lay audiences' assessment of expert knowledge. The source of information can serve as a heuristic cue in evaluating content quality [Ismagilova et al., 2020]. Transferred to AI-generated information, we posit that the labeling of the source influences how the strength and correctness of the arguments are assessed. A more favorable attitude toward a source is assumed to result in a higher rating of the strength and correctness of the argument.

Based on studies suggesting a high level of trust and ascribed expertise to medical experts [e.g., Wissenschaft im Dialog, 2023], the expert label is assumed to lead to rather positive quality assessments. For AI, studies show that the public is more skeptical [e.g., Wissenschaft im Dialog, 2023], and the AI-labeling can lead to lower quality assessments [Karinshak et al., 2023; Lim & Schmälzle, 2024; Reis et al., 2024]. Since our first study indicated lower ratings for AI-generated arguments, we hypothesize that source labeling will further amplify these differences. We further extend research by comparing labeled information to unlabeled information [Teigen et al., 2024], assuming source heuristics enhance quality assessments of expert-authored arguments and lower those of AI-generated arguments in comparison with unlabeled arguments. This leads to the following hypotheses:

> **H1a-c:** (a) The argument strength of the argument with the expert label is rated as the strongest. (b) The rating of the expert label is followed by the argument without a label. (c) The argument with the AI label is rated as the weakest.
> **H2a-c:** (a) The correctness of the argument with the expert label is rated the highest. (b) The rating of the expert label is followed by the argument without a label. (c) The correctness of the argument with the AI label is rated lowest.

Given calls for transparency in AI-assisted content creation [Deutsche Forschungsgemeinschaft, 2023], we included a "collaboration" label indicating expert content created with AI assistance. The effect of collaboration labels on the perception of an argument has rarely been investigated. Therefore, a mitigating or reinforcing effect in comparison to the expert label seems possible. Initial evidence of Reis et al. [2024] showed that collaboration labels are perceived as less reliable than labels referring to physicians but more reliable than AI alone. Against this backdrop, we assume that, compared to established public medical experts, the collaboration label will lead to lower perceptions of argument strength and correctness. Therefore, we derived the following hypotheses:

> **H3:** The argument strength is rated stronger for the argument with the expert label than for the argument with the collaboration label.
> **H4:** Correctness is rated higher for the argument with the expert label than for the argument with the collaboration label.

Compared to the AI label, the reference to experts might result in better assessments of the collaboration label, which aligns with findings from Lim and Schmälzle [2024] revealing that AI arguments created with expert prompts receive better ratings. However, findings of Reis et al. [2024] contradict this assumption. Their findings did not reveal a difference between

the collaboration and the AI labels in the context of human physicians. Nevertheless, since we focus on established public medical experts instead of an unknown single physician, we assume that the positive attitudes toward the experts might be stronger. Therefore, we derived the following hypotheses:

H5: The argument strength is rated stronger for the argument with the collaboration label than for the argument with the AI label.
H6: Correctness is rated higher for the argument with the collaboration label than for the argument with the AI label.

3.1 ▪ *Moderating effects of trust and innovativeness*

To provide further insights into the underlying mechanisms of the effects of source disclosure, we focus on two moderators that are central to understanding variations in quality assessments: trust and innovativeness. Both have been identified as key factors in prior research on AI-generated information [Jung et al., 2017; Karinshak et al., 2023; Khan et al., 2019]. By focusing on trust and innovativeness, we aim to capture two complementary dimensions underlying quality assessments: attitudes toward the source providing the arguments (trust) and personality traits of potential users (innovativeness).

Trust was considered to highlight its heuristic function in source credibility. Trust assessments underlying source credibility play a central role in heuristic information processing, where mental shortcuts shape attitudes toward information and its source [Cummings, 2014]. It is defined as one's willingness to be vulnerable and assign responsibility to the object of trust [e.g., Mayer et al., 1995; Sztompka, 1999]. Higher trust in sources affects a more favorable perception of the value of (health) information [Link, 2019]. In the context of AI-generated versus expert-generated content, it has been suggested that trust in AI can drive differences in perceived quality [Jung et al., 2017; Karinshak et al., 2023]. Specifically, individuals with higher trust in AI may rely on this heuristic to form more positive evaluations of AI-labeled arguments, reducing the differences between AI and expert labels (H7–H8). Conversely, trust in science or established medical institutions like the STIKO might amplify the positive evaluations of arguments with expert (H9a, H10a, H11a, and H12a) or collaboration labels (H9b, H10b, H11b, and H12b) compared to the AI label due to their alignment with the trusted source. The more positive evaluation is assumed to result in higher differences. These considerations lead to the following hypotheses:

H7: The higher the trust in AI, the smaller the differences in the assessment of argument strength between the argument with the AI label and (a) expert label, (b) no label, and (c) collaboration label.
H8: The higher the trust in AI, the smaller the differences in the assessment of correctness between the argument with the AI label and (a) expert label, (b) no label, and (c) collaboration label.
H9: The higher the trust in science, the larger the differences in the assessment of argument strength between the argument with the AI label and (a) expert label, and (b) collaboration label.

H10: The higher the trust in science, the larger the differences in the assessment of correctness between the argument with the AI label and (a) expert label, and (b) collaboration label.
H11: The higher the trust in the STIKO, the larger the differences in the assessment of (a) argument strength between the argument with the AI label and (a) expert label, and (b) collaboration label.
H12: The higher the trust in the STIKO, the larger the differences in the assessment of correctness between the argument with the AI label and (a) expert label, and (b) collaboration label.

A second moderator considered is an individual's innovativeness, which is a personality trait indicating the extent to which an individual is inclined to use new technologies [Goldsmith, 2011; Goldsmith & Hofacker, 1991]. Transferred to AI-generated information, initial research showed that more innovative individuals experience less psychological distance to AI, evaluate recommendations by AI more positively [Jang et al., 2023], and rate its quality more positively [Khan et al., 2019]. Moreover, Jung et al. [2017] proposed that the exceptionally high average affinity for new technologies may explain a more favorable evaluation of news articles written by algorithms. This suggests that innovativeness may act as a buffer against bias toward AI-labeled arguments, resulting in smaller differences in quality assessments between AI-labeled arguments and those with other labels. Based on these findings, we propose that innovativeness is a moderator of the quality assessment of AI-generated information, leading to the following hypotheses:

H13: The higher an individual's innovativeness, the smaller the differences in the assessment of argument strength between the argument with the AI label and (a) expert label, (b) no label, and (c) collaboration label.
H14: The higher an individual's innovativeness, the smaller the differences in the assessment of correctness between the argument with the AI label and (a) expert label, (b) no label, and (c) collaboration label.

3.2 ▪ *Methods*

The second study was preregistered (https://aspredicted.org/5sfg-ncyy.pdf). The examination of the collaboration label and the influence of trust in science and the STIKO was added to the preregistration.

3.2.1 ▪ *Participants and procedure*

We conducted a between-person online experiment. The participants were recruited via a German online access panel [SoSci-Panel; Leiner, 2016]. The final sample consisted of N = 1,029 participants after excluding 130 individuals who did not answer the manipulation check correctly, 50 who completed the survey too quickly [Leiner, 2019], and two who did not answer the dependent variables. On average, the participants were 52.1 years old (SD = 16.1), and 61.2% identified as female. They were highly educated, with 82.4% having a high school diploma, 13.8% having a secondary school diploma, and 3.2% having a lower secondary school diploma or no high school diploma.

Again, this type of data collection was considered exempt from the need for ethical approval according to German legal regulations. To meet ethical guidelines, participants were informed about the voluntary nature of participation, their right to withdraw, and anonymity, as well as how data would be collected, processed, and stored. Further, they were asked to give their informed consent to participate. In addition, after reading the stimulus and answering the question, they were provided with information about the study's objectives, the true source of the stimulus material, and were asked again for their informed consent.

3.2.2 ▪ *Stimulus*

The stimulus was one argument from the first study, which was varied using different labels. We selected the argument "Reduction of individual risks" from experts, which had consistently received rather positive ratings in the first study, to increase the probability that possible differences could be attributed to the label. The labels used were (1) ChatGPT, (2) STIKO, (3) STIKO with ChatGPT, and (4) no label. The labels were displayed below the text like a reference (see appendix A, Figure 1). Participants were also shown a short explanatory text explaining that the label was the author of the argument and what ChatGPT or STIKO are.

3.2.3 ▪ *Measures*

Argument strength and correctness. Perceived argument strength [Zhao et al., 2011] and correctness [Kohring & Matthes, 2004] were measured in line with the first study. Based on satisfactory reliability scores (Argument strength: α = .91–.93; Correctness: α = .83–.86), mean indices were calculated (see Table 2).

Trust. Trust in various objects was measured with single items asking how much the participants trusted science (M = 4.19, SD = 0.80), the STIKO (M = 3.62, SD = 1.16), and AI (M = 2.32, SD = 0.93), which could be answered on 5-point Likert-type scales (see appendix A, Table 3). The decision to use single-item measures was based on prior research [e.g., Castro et al., 2023; Reif & Guenther, 2021], which suggests that single items can be a valid compromise when assessing broad constructs across multiple domains within time-constrained survey designs.

Innovativeness. Based on Goldsmith [2011], we measured an individual's innovativeness in the context of AI by adapting a seven-item scale to the context under investigation. The items were answered on a 5-point Likert-type scale. As the internal consistency was satisfactory, we calculated a mean index (α = .86; M = 3.14, SD = 0.93). We reversed the scale so that higher values indicate higher innovativeness.

3.2.4 ▪ *Manipulation check*

The manipulation check consisted of one question asking the participants for the source of the presented argument after measuring the dependent variables, providing four possible answers (1 = ChatGPT, 2 = STIKO, 3 = STIKO with ChatGPT, 4= no information). Participants could also indicate that they could not remember.

3.2.5 ▪ *Data analysis*

To test hypotheses H1–H6, ANOVAs were conducted using SPSS 27 with the labeling of the source as independent and perceived argument strength and correctness as dependent variables. As homogeneity of variances was given, post hoc tests were conducted with Bonferroni. For testing H7–H14, moderation analyses were conducted using the PROCESS macro by Hayes [2022]. The labeling of the source was the independent variable, while trust in AI (H7–H8), science (H9–H10) or STIKO (H11–H12), and innovativeness (H13–H14) served as moderators. Perceived argument strength and correctness were the dependent variables.

3.3 ▪ *Results*

3.3.1 ▪ *Effects of labeling*

Regarding H1a-c, which addressed the argument strength of the various labels, the ANOVA revealed that the level of perceived argument strength ($F(3, 1025) = 8.003$, $p < .001$, $\eta^2 = .02$) differed significantly between the different labels. However, the effect was rather small. In line with the hypotheses, the argument with the expert label was rated stronger than the argument without a label (cf. H1a/b), which in turn was rated stronger than the argument with the AI label (cf. H1c) (see Table 2). However, only the differences between the expert label and AI label (.40, 95%-CI[.18, .62], $p < .001$), and between no label and AI label (.22, 95%-CI[.003, .45], $p = .04$) were significant. Therefore, H1 was partly supported.

Regarding perceived correctness addressed in H2a-c, the ANOVA showed that the perceived correctness of the argument ($F(3, 1026) = 18.613$, $p < .001$, $\eta^2 = .052$) differed significantly between the different labels. Although the effect was rather small, the argument with the expert label was rated the highest, followed by the argument without a label and the AI label (see Table 2). The differences between the expert label and AI label (.40, 95%-CI[.18, .62], $p < .001$), between expert label and no label (.40, 95%-CI[.18, .62], $p < .001$), and between no label and AI label (.40, 95%-CI[.18, .62], $p < .001$) were significant, thus supporting H2a-c.

Table 2. Descriptive results of argument strength and correctness for different authorship labels.

Label	**ChatGPT** (n = 264–265)	**STIKO** (n = 265)	**STIKO with ChatGPT** (n = 243–244)	**No label** (n = 255)
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Perceived argument strength	3.17^{ab} (0.93)	3.57^{ac} (1.00)	3.32^{c} (0.96)	3.40^{b} (0.92)
Perceived correctness	3.59^{d} (0.90)	4.13^{de} (0.84)	3.70^{e} (0.89)	3.83^{d} (0.88)

N = 1,029, Results of two ANOVAs; superscript letters indicate significant differences, values sharing the same letter differ significantly; Perceived argument strength: $F(3, 1025) = 8.003$, $p < .001$, $\eta^2 = .02$; Perceived correctness: $F(3, 1026) = 18.613$, $p < .001$, $\eta^2 = .052$.

H3, H4, H5, and H6 examined the effect of the collaboration label in comparison to the expert or AI label. Regarding H3 and H4, we found that the perceived strength and correctness of the argument with the collaboration label were lower than those of the argument with the expert label. Both differences were significant (Argument strength: -.25, 95%-CI[-.47, -.03], $p = .02$; Correctness: -.43, 95%-CI[-.63, -.22], $p < .001$). Therefore, H3 and H4 were supported.

The differences between the AI and collaboration label were neither significant for argument strength nor for correctness (see Table 2). Thus, H5 and H6 were not supported.

3.3.2 ▪ *Moderating effects of trust*

Focusing on the postulated moderation effects, we proposed that higher trust in AI reduces the differences in perceived argument strength (H7a) and correctness (H8a) between the expert label and the AI label. Although the overall models were significant (Argument strength: R^2 = .069, F(3, 526) = 12.95, p < .001, Correctness: R^2 = .111, F(3, 525) = 21.75, p < .001) trust in AI did not moderate the effect between labeling and perceived argument strength (ΔR^2 = 0.00%, F(1, 526) = 0.09, p = .759, 95% CI[-0.149, 0.205]) nor between labeling and perceived correctness (ΔR^2 = 0.00%, F(1, 525) = 0.01, p = .914, 95% CI[-0.153, 0.171]). Thus, H7a and H8a were not supported.

For the comparison of no label and the AI label (H7b and H8b), the overall models were significant (Argument strength: R^2 = .049, F(3, 516) = 8.84, p < .001, Correctness: , R^2 = .039, F(3, 515) = 7.00, p < .001), but trust in AI again did not moderate the relationship between labeling and perceived argument strength nor correctness (Argument strength: ΔR^2 = 0.00%, F(1, 516) = 0.015, p = .902, 95% CI[-0.091, 0.080], Correctness: ΔR^2 = 0.00%, F(1, 515) = 0.047, p = .828, 95% CI[-0.093, 0.075]). Therefore, H7b and H8b were not supported.

Comparing the collaboration and AI label (H7c and H8c), the results revealed that trust in AI did not serve as a moderator (Argument strength: ΔR^2 = 0.00%, F(1, 505) = 0.003, p = .957, 95% CI[-0.173, 0.183], Correctness: ΔR^2 = 0.00%, F(1, 503) = 0.013, p = .911, 95% CI[-0.181, 0.162]) for effects of labeling on perceived argument strength and correctness (Argument strength: R^2 = .041, F(3, 505) = 7.21, p < .001, Correctness: , R^2 = .026, F(3, 503) = 4.54, p < .01), leading to H7c and H8c not being supported.

Regarding trust in science, we proposed that higher trust in science increases the differences in perceived argument strength (H9) and correctness (H10) between (a) expert and AI label and between (b) collaboration label and AI label. Comparing the expert and AI label, the overall models were significant (Argument strength: R^2 = .173, F(3, 526) = 36.54, p < .001, Correctness: R^2 = .233, F(3, 525) = 53.22, p < .001). Trust in science moderated the effect between labeling and perceived argument strength (ΔR^2 = .013, F(1, 526) = 8.29, p < .01, 95% CI[-0.874, -0.088]) and between labeling and correctness (ΔR^2 = .015, F(1, 525) = 10.37, p < .01, 95% CI[-0.444, - 0.108]). In line with H9a and H10a, higher trust in science increased the differences in assessments of argument strength and correctness between the AI and expert labels.

Comparing the collaboration label and AI label (H9b and H10b), the overall models were significant (Argument strength: R^2 = .081, F(3, 505) = 14.90, p < .001, Correctness: R^2 = .086, F(3, 503) = 15.82, p < .001). However, trust in science did not moderate the effect between labeling and perceived argument strength (ΔR^2 = .005, F(1, 505) = 2.82, p = .094, 95% CI[-0.394, 0.031]) nor between labeling and perceived correctness (ΔR^2 = .007, F(1, 503) = 3.68, p = .056, 95% CI[-0.397, 0.005]). Therefore, H9b and H10b were not supported.

H11 and H12 addressed the role of trust in the STIKO. Comparing the expert and AI label (H11a and H12a), the overall models were significant (Argument strength: R^2 = .280, F(3, 526) = 68.33, p < .001, Correctness: R^2 = .277, F(3, 525) = 66.99, p < .001). Trust in the STIKO

significantly moderated the effect between labeling and perceived argument strength (ΔR^2 = .015, *F*(1, 526) = 10.67, *p* < .01, 95% CI[-0.320, -0.080]) and between labeling and perceived correctness (ΔR^2 = .011, *F*(1, 525) = 8.03, *p* < .01, 95% CI[-0.273, -0.049]). Higher trust in the STIKO increased the difference in assessments of argument strength and correctness between the AI and expert labels, supporting H11a and H12a.

The overall models for the comparison of collaboration and AI label (H11b and H12b) were also significant (Argument strength: R^2 = .234, *F*(3, 505) = 51.50, *p* < .001, Correctness: R^2 = .157, *F*(3, 503) = 31.13, *p* < .001). Trust in the STIKO moderated the effect of labeling on perceived argument strength (ΔR^2 = .014, *F*(1, 505) = 9.295, *p* < .01, 95% CI[-0.323, -0.070]). Higher trust in the STIKO increased the difference in the assessments of argument strength, which is in line with H11b. In contrast, we found no moderating effect on perceived correctness (ΔR^2 = .006, *F*(1, 503) = 3.715, *p* = .055, 95% CI[-0.249, 0.002]). Therefore, H12b was not supported.

3.3.3 ▪ *Moderating effects of innovativeness*

Further, we proposed that an individual's innovativeness moderates the effect of labeling on the perception of argument strength (H13) and correctness (H14) between the (a) expert and AI label, (b) no label and AI label, and (c) collaboration label and AI label. Comparing the expert label and the AI label, the overall models were significant (Argument strength: R^2 = .044, *F*(3, 523) = 7.98, *p* < .001; Correctness: R^2 = .097, *F*(3, 522) = 18.64, *p* < .001), but the moderation analyses revealed that innovativeness did neither moderate the effect between labeling and perceived argument strength (ΔR^2 = .000, *F*(1, 523) = 0.109, *p* = .741, 95% CI[-0.212, 0.151]) nor between labeling and correctness (ΔR^2 = .000, *F*(1, 522) = 0.07, *p* = .798, 95% CI[-0.185, 0.143]). Therefore, H13a and H14a were not supported.

Regarding the comparison of the argument without a label and the argument with AI label, the findings revealed that the overall models were significant (Argument strength: R^2 = .018, *F*(3, 511) = 3.10, *p* < .05, Correctness: , R^2 = .023, *F*(3, 510) = 3.92, *p* < .01), but no moderating effects were found (Argument strength: ΔR^2 = .000, *F*(1, 511) = 0.077, *p* = .781, 95% CI[-0.074, 0.098]; Correctness: ΔR^2 = .002, *F*(1, 510) = 0.918, *p* = .338, 95% CI[-0.085, 0.248]). Thus, H13b and H14b were not supported.

Furthermore, an individual's innovativeness did not moderate the effects of labeling on perceived argument strength (ΔR^2 = .000, *F*(1, 502) = 0.175, *p* = .676, 95% CI[-0.213, 0.138]) or perceived correctness (ΔR^2 = .002, *F*(1, 500) = 1.037, *p* = .309, 95% CI[-0.252, 0.080]) in the comparison of the collaboration label and AI label (Argument strength: R^2 = .008, *F*(3, 502) = 7.21, *p* =.28, Correctness: , R^2 = .007, *F*(3, 500) = 1.24, *p* = .293). Thus, H13c and H14c were not supported.

3.4 ▪ *Discussion of the results of study 2*

The first objective of our second study was to examine how authorship labels influence recipients' quality perceptions of arguments regarding influenza vaccination. In line with previous research [Karinshak et al., 2023; Lim & Schmälzle, 2024; Teigen et al., 2024], labeling arguments as human-generated (by medical experts) led to higher quality assessments than labeling them as AI-generated for both argument strength and

correctness. The AI label resulted in lower quality perceptions than no label. Only in comparison with the collaboration label, the AI label was not rated significantly worse. However, the collaboration label was assessed significantly lower than the expert label, suggesting that AI involvement alone reduces perceived argument strength and correctness of arguments on the influenza vaccination.

A second objective was to extend the current state of research regarding possible explanations for the different perceptions of the labels by investigating the moderating roles of trust and innovativeness. Our analyses revealed that trust in AI did not moderate differences in the quality perception between the AI label and other labels, contradicting the findings of Karinshak et al. [2023], who found moderating effects of a lack of trust in AI on the perception of labeling between expert and AI labels in a vaccination context. This discrepancy may be due to lower levels of trust in AI and less variance in our sample. Approximately half of our participants reported that they had never used AI, such as ChatGPT. This could lead to an inadequate understanding of ChatGPT, a lack of experience to build trust, or less stable trust assessments toward AI [Barber, 1983], which might not serve as heuristic cues.

The German context may also play a role, as the comparatively high level of access to medical professionals in Germany makes the role of doctors in providing health information arguably more important than that of digital sources [Link et al., 2022]. It is therefore possible that German residents are more skeptical about the use of AI for health purposes. Additionally, the measurement of trust might impact the absence of the moderation effect. Besides the single item, trust in AI was measured very broadly instead of focusing specifically on ChatGPT as the source of the argument provided.

Contrary to previous research [Jang et al., 2023; Jung et al., 2017; Khan et al., 2019], innovativeness did not moderate the effect of labeling on quality assessments. More innovative individuals did not assess AI-generated information as being of higher quality, leading to smaller differences between the AI label and other labels. One possible explanation could lie in the health context of the study, which is more sensitive than AI-generated recommendations in other contexts [Jang et al., 2023; Khan et al., 2019], and may require more positive attitudes toward AI than individuals' innovativeness, which indicates openness to new technologies. This aligns with findings by Teigen et al. [2024], which suggest that labeling effects vary across domains such as health, politics, or finance.

Trust in science and the STIKO moderated the effect of labeling. In line with H9a, H10a, H11a, and H12a, high trust in science or the STIKO increased the differences between the quality assessments of the expert and AI labels. Even at lower trust in science or the STIKO, expert-labeled arguments were perceived as more correct than AI-labeled ones, and with low trust in the STIKO, this was also the case for argument strength. The effect increased with increasing trust. When comparing the collaboration label to the AI label, high trust in the STIKO led recipients to perceive the argument with the collaboration label as significantly stronger than with the AI label. This indicates that collaboration alone is not sufficient to result in a better quality assessment; it is necessary that the experts involved are perceived as trustworthy. Additionally, it seems important that the contributing experts are specifically trusted, as evidenced by the finding that trust in science in general did not lead to any differences between AI and collaboration labels.

4 ▪ Discussion

Our two studies aimed to investigate whether AI can provide adequate information from the recipient's perspective and how the information provided is perceived when authorship is unknown (Study 1) or disclosed (Study 2).

4.1 ▪ *Key findings*

Overall, our two-study design revealed that both AI and experts were able to provide convincing and correct information about influenza vaccination. However, unlike previous research [Jakesch et al., 2019; Karinshak et al., 2023; Lim & Schmälzle, 2023; Palmer & Spirling, 2023], our results showed that recipients perceive experts' arguments as stronger and more correct, even when authorship is unknown. We attribute this to our approach of simulating natural usage without pre-training ChatGPT. Notably, the perceived quality of experts' arguments was more consistent, while the perceived quality of ChatGPT's arguments exhibited greater variation. This suggests that when users search for vaccine-related information using ChatGPT, they may encounter information they find less appropriate than when they seek information from experts.

Furthermore, the results of our second study showed that labeling information as AI-generated consistently led to lower quality assessments compared to expert-labeled or unlabeled arguments, reinforcing the results from Study 1. The use of ChatGPT by experts was also perceived more negatively than information labeled solely as authored by experts.

Aiming to identify the reasons for the poorer assessment of AI-generated arguments, our experimental study indicated that the already better evaluation of information from medical experts increased even further when people trusted science or the STIKO. However, the poorer assessment of AI-generated arguments could not be compensated by a higher level of trust in AI, which contradicts previous research [Karinshak et al., 2023]. Additionally, a high level of innovativeness did not affect quality perceptions, likely due to the health context, where arguments are rated more systematically.

4.2 ▪ *Limitations*

While our studies contribute to the understanding of individuals' perceptions of AI-generated health information, several limitations should be considered and can guide future research. First, our sample was highly educated and older than the general German population. While education may be linked to higher trust in science [Wissenschaft im Dialog, 2023], age differences could also impact attitudes toward AI. Future research should explore how both factors influence the perception of AI-generated health information. Second, we measured trust in AI in a generalized manner, whereas the AI label specifically displayed ChatGPT as the source. Future research should consider measuring trust more specifically. Third, our focus on influenza vaccination limits generalizability to other health contexts, warranting further research. Fourth, we examined only a subset of possible moderators; future studies should explore additional factors such as attitudes toward AI, including the machine heuristic [Sundar, 2008, p. 83], the tendency to view machine decisions as "objective" and "free from ideological bias." Another promising moderator could be an individual's involvement, given its importance in dual process models [Petty et al., 1981]. Fifth, our study used the GPT-3

version of ChatGPT, available for free during data collection. Since GPT-4 surpasses GPT-3 in several abilities [Ahsan et al., 2023], the results of our first study might differ with GPT-4. Additionally, as ChatGPT is not specialized in medical contexts [Zhou et al., 2023], future research could compare health-related AIs with medical experts. Sixth, since we focused on recipients' perceptions, we cannot determine whether expert arguments were inherently superior. Future studies should investigate the reasons for the different perceptions when the source is unknown. Seventh, because we used only one argument as a stimulus in the second study, we cannot ascertain whether the actual strength of the argument might influence the label effect. Therefore, future research should compare weak and strong arguments. Lastly, our study did not consider the interactive capabilities of chatbots, which are designed for dialogue. Since dialogic engagement has been shown to effectively counter misinformation [Costello et al., 2024], future research should explore the potential impact of interactive chatbot features on the perception of AI-generated health information.

5 ▪ Conclusion

Regarding the potential of AI for providing health information [Deiana et al., 2023; Lee et al., 2023; Meng et al., 2024], our results indicate that AI can provide users with information perceived as high quality. However, this potential is limited by varying perceptions of content quality. While AI allows users to ask personalized follow-up questions [Lee et al., 2023] and gain detailed information that can contribute to a convincing and correct overall picture of a health issue, adequate health literacy is still essential to assess the quality of the information and evaluate its helpfulness or applicability. Moreover, AI-generated content depends on how questions are formulated, posing a risk of misleading answers if prompts are unclear [Deiana et al., 2023]. This can be particularly problematic in the context of health information. Future efforts should therefore focus on equipping users with the necessary skills to effectively utilize AI for health information seeking.

Additionally, our findings show that the potential of AI might also fail because individuals are skeptical toward AI and attribute less quality to arguments if they know that AI was involved in their generation. Efforts are needed to support individuals in making informed trust assessments and managing risk perceptions toward AI. Given the unclear sources of AI training data and the prevalence of misinformation online [e.g., Deiana et al., 2023], as well as the fact that AI chatbots are probabilistic language models and do not provide verified knowledge [Yildirim & Paul, 2024], it may be beneficial for individuals to rate expert information as higher quality. As AI continues to expand across various fields, labeling expert-generated information could be advantageous. Furthermore, fostering trust in scientific and medical institutions like the STIKO remains crucial, as such trust seems to enhance the perceived quality of their information.

A ▪ Details on the study design

Table 3. Overview of the measures.

Construct	Examples of item wording	Response scale	Descriptive	Internal Consistency	Source
Perceived Argument Strength	The statement is a reason for getting the influenza vaccination that is convincing. The statement gives a reason for getting the influenza vaccination that is important to me.	five-point Likert-type scale from 1 "Strongly disagree" to 5 "Strongly agree"; 1 "Not at all" to 5 "Completely"; 1 "Very weak" to 5 "Very strong"	see Table 1 and Table 2	Study 1: Cronbach's α = .92 - .95 Study 2: Cronbach's α = .91 - .93	Zhao et al. [2011]
Correctness	The text presents the facts as they are. The information given is true.	five-point Likert-type scale from 1 "Strongly disagree" to 5 "Strongly agree"	see Table 1 and Table 2	Study 1: Cronbach's α = .92 - .95 Study 2: Cronbach's α = .83 - .86	Kohring and Matthes [2004]
Trust	To what extent do you trust the following institutions or technologies?	five-point Likert-type scale from 1 "Not at all" to 5 "Completely"			Self-developed
	Science		*M* = 4.19 *SD* = 0.80		
	Standing Committee on Vaccination		*M* = 3.62 *SD* = 1.16		
	AI		*M* = 2.32 *SD* = 0.93		
Innovativeness	I am suspicious of new inventions and ways of thinking related to artificial intelligence. (reversed)	five-point Likert-type scale from 1 "Strongly disagree" to 5 "Strongly agree"	*M* = 3.14 *SD* = 0.93	Cronbach's α = .86	Goldsmith [2011], adapted to the context of AI

Die echte Grippe (Influenza) ist keine einfache Erkältungskrankheit ("grippaler Infekt"), sondern eine ernstzunehmende Infektion, die durch Influenzaviren verursacht wird. Nach einer Ansteckung mit Influenzaviren beginnt bei etwa einem Drittel der Betroffenen eine Grippe plötzlich mit hohem Fieber (über 38,5 °C), trockenem Husten, Kopf-, Hals-, Muskel- und Gliederschmerzen, Abgeschlagenheit und manchmal Übelkeit/Erbrechen sowie Schweißausbrüchen. Die Grippeimpfung beugt dem vor.

(Quelle: Ständige Impfkommission und Chat GPT)

Figure 1. Example of the stimulus material.

English translation. The real flu (influenza) is not a simple cold (“flu-like infection”) but a serious infection caused by influenza viruses. After contracting influenza viruses, about one-third of those affected experience a sudden onset of flu symptoms, including high fever (above 38.5°C), dry cough, headaches, sore throat, muscle and joint pain, fatigue, and sometimes nausea/vomiting as well as sweating episodes. The flu vaccination helps prevent this.

References

Ahsan, M. M. T., Rahaman, M. S., & Anjum, N. (2023). From ChatGPT-3 to GPT-4: a significant leap in AI-driven NLP tools. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4404397

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a Public Social Media Forum. *JAMA Internal Medicine*, *183*, 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Barber, B. (1983). *The logic and limits of trust*. Rutgers University Press. http://hdl.handle.net/10822/796491

Brettschneider, F., & Bachl, M. (2020). *Umfrage zur Oberbürgermeisterwahl in Stuttgart* [Survey on the mayoral election in stuttgart]. Universität Hohenheim. https://komm.uni-hohenheim.de/uploads/media/2020-11_OB-Wahl_Stuttgart_Welle_1.pdf

Carcioppolo, N., Yang, F., & Yang, Q. (2016). Reducing, maintaining, or escalating uncertainty? The development and validation of four uncertainty preference scales related to cancer information seeking and avoidance. *Journal of Health Communication*, *21*, 979–988. https://doi.org/10.1080/10810730.2016.1184357

Castro, M. S., Bahli, B., Ferreira, J. J., & Figueiredo, R. (2023). Comparing single-item and multi-item trust scales: insights for assessing trust in project leaders. *Behavioral Sciences*, *13*, 786. https://doi.org/10.3390/bs13090786

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*. https://doi.org/10.1126/science.adq1814

Cummings, L. (2014). The “trust” heuristic: arguments from authority in public health. *Health Communication*, *29*, 1043–1056. https://doi.org/10.1080/10410236.2013.831685

Deiana, G., Dettori, M., Arghittu, A., Azara, A., Gabutti, G., & Castiglia, P. (2023). Artificial intelligence and public health: evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines*, *11*, 1217. https://doi.org/10.3390/vaccines11071217

Deutsche Forschungsgemeinschaft. (2023). *Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das förderhandeln der DFG* [Statement by the executive committee of the german research foundation (dfg) on the influence of generative models for text and image production on the sciences and the dfg's funding activities]. https://www.dfg.de/resource/blob/289674/ff57cf46c5ca109cb18533b21fba49bd/230921-stellungnahme-praesidium-ki-ai-data.pdf

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Goldsmith, R. E. (2011). The validity of a scale to measure global innovativeness. *Journal of Applied Business Research (JABR)*, *7*, 89. https://doi.org/10.19030/jabr.v7i2.6249

Goldsmith, R. E., & Hofacker, C. F. (1991). Measuring consumer innovativeness. *Journal of the Academy of Marketing Science*, *19*, 209–221. https://doi.org/10.1007/bf02726497

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' perception of computer-generated news: credibility, expertise and readability. *Journalism*, *19*, 595–610. https://doi.org/10.1177/1464884916641269

Hayes, A. F. (2022). Methodology in the social sciences. In A. F. Hayes (Ed.), *Introduction to mediation, moderation and conditional process analysis: a regression-based approach* (3rd ed.). The Guilford Press.

Hershenhouse, J. S., Mokhtar, D., Eppler, M. B., Rodler, S., Storino Ramacciotti, L., Ganjavi, C., Hom, B., Davis, R. J., Tran, J., Russo, G. I., Cocci, A., Abreu, A., Gill, I., Desai, M., & Cacciamani, G. E. (2024). Accuracy, readability and understandability of large language models for prostate cancer information to the public [Advance online publication]. *Prostate Cancer and Prostatic Diseases*. https://doi.org/10.1038/s41391-024-00826-y

Ismagilova, E., Slade, E., Rana, N. P., & Dwivedi, Y. K. (2020). The effect of characteristics of source credibility on consumer behaviour: a meta-analysis. *Journal of Retailing and Consumer Services*, *53*, 101736. https://doi.org/10.1016/j.jretconser.2019.01.005

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: how the perception that profile text was written by AI affects trustworthiness. In S. Brewster, G. Fitzpatrick, A. Cox & V. Kostakos (Eds.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300469

Jang, W., Chang, Y., Kim, B., Lee, Y. J., & Kim, S.-C. (2023). Influence of personal innovativeness and different sequences of data presentation on evaluations of explainable artificial intelligence. *International Journal of Human-Computer Interaction*, *40*, 4215–4226. https://doi.org/10.1080/10447318.2023.2209995

Johnson, J. D., & Meischke, H. (1993). A comprehensive model of cancer-related information seeking applied to magazines. *Human Communication Research*, *19*, 343–367. https://doi.org/10.1111/j.1468-2958.1993.tb00305.x

Jung, J., Song, H., Kim, Y., Im, H., & Oh, S. (2017). Intrusion of software robots into journalism: the public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in Human Behavior*, *71*, 291–298. https://doi.org/10.1016/j.chb.2017.02.022

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 1–29. https://doi.org/10.1145/3579592

Khan, A., Masrek, M. N., & Mahmood, K. (2019). The relationship of personal innovativeness, quality of digital resources and generic usability with users' satisfaction: a Pakistani perspective. *Digital Library Perspectives*, *35*, 15–30. https://doi.org/10.1108/dlp-12-2017-0046

Kohring, M., & Matthes, J. (2004). Revision und Validierung einer Skala zur Erfassung von Vertrauen in Journalismus [Revision and validation of a scale to measure trust in journalism]. *Medien & Kommunikationswissenschaft*, *52*, 377–385. https://doi.org/10.5771/1615-634x-2004-3-377

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits and risks of GPT-4 as an AI chatbot for medicine (J. M. Drazen, I. S. Kohane & T.-Y. Leong, Eds.). *New England Journal of Medicine*, *388*, 1233–1239. https://doi.org/10.1056/nejmsr2214184

Leiner, D. J. (2016). Our research's breadth lives on convenience samples: a case study of the online respondent pool "SoSci Panel". *Studies in Communication | Media*, *5*, 367–396. https://doi.org/10.5771/2192-4007-2016-4-367

Leiner, D. J. (2019). Too fast, too straight, too weird: non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*, 229–248. https://doi.org/10.18148/SRM/2019.V13I3.7403

Lim, S., & Schmälzle, R. (2023). Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. *Frontiers in Communication*, *8*. https://doi.org/10.3389/fcomm.2023.1129082

Lim, S., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages. *Computers in Human Behavior: Artificial Humans*, *2*, 100058. https://doi.org/10.1016/j.chbah.2024.100058

Link, E. (2019). *Vertrauen und die Suche nach Gesundheitsinformationen: Eine empirische Untersuchung des Informationshandelns von Gesunden und Erkrankten* [Trust and health information seeking: An empirical investigation of information seeking behavior of healthy and diseased individuals]. Springer VS. https://doi.org/10.1007/978-3-658-24911-3

Link, E., Baumann, E., Kreps, G. L., Czerwinski, F., Rosset, M., & Suhr, R. (2022). Expanding the health information national trends survey research program internationally to examine global health communication trends: comparing health information seeking behaviors in the U.S. and Germany. *Journal of Health Communication*, *27*, 545–554. https://doi.org/10.1080/10810730.2022.2134522

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*, 709. https://doi.org/10.2307/258792

Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., Zhang, M., Cao, C., Wang, J., Wang, X., Gao, J., Wang, Y.-G.-S., Ji, J.-m., Qiu, Z., Li, M., Qian, C., Guo, T., Ma, S., Wang, Z., … Tang, Y.-D. (2024). The application of large language models in medicine: a scoping review. *iScience*, *27*, 109713. https://doi.org/10.1016/j.isci.2024.109713

Palmer, A., & Spirling, A. (2023). Large Language Models can argue in convincing ways about politics, but humans dislike AI authors: implications for governance. *Political Science*, *75*, 281–291. https://doi.org/10.1080/00323187.2024.2335471

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, *41*, 847–855. https://doi.org/10.1037/0022-3514.41.5.847

Reif, A., & Guenther, L. (2021). How representative surveys measure public (dis)trust in science: a systematisation and analysis of survey items and open-ended questions. *Journal of Trust Research*, *11*, 94–118. https://doi.org/10.1080/21515581.2022.2075373

Reis, M., Reis, F., & Kunde, W. (2024). Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine*, *30*, 3098–3100. https://doi.org/10.1038/s41591-024-03180-7

Robert Koch Institute. (2023). *Epidemiologisches Bulletin: Empfehlungen der Ständigen Impfkommission beim Robert Koch-Institut 2023* [Epidemiological bulletin: Recommendations of the standing committee on vaccination at the robert koch institute 2023]. https://edoc.rki.de/bitstream/handle/176904/10636/EB-4-2023-Deutsch.pdf

Song, H., Xia, Y., Luo, Z., Liu, H., Song, Y., Zeng, X., Li, T., Zhong, G., Li, J., Chen, M., Zhang, G., & Xiao, B. (2023). Evaluating the performance of different Large Language Models on health consultation and patient education in Urolithiasis. *Journal of Medical Systems*, *47*, 125. https://doi.org/10.1007/s10916-023-02021-3

Sundar, S. S. (2008). The MAIN model: a heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *The John D. and Catherine T. MacArthur Foundation series on digital media and learning* (pp. 73–100). The MIT Press.

Sztompka, P. (1999). *Trust: a sociological theory*. Cambridge University Press.

Teigen, C., Madsen, J. K., George, N. L., & Yousefi, S. (2024). Persuasiveness of arguments with AI-source labels. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*. https://escholarship.org/uc/item/6t82g70v

Trepte, S., Baumann, E., Hautzinger, N., & Siegert, G. (2005). Qualität gesundheitsbezogener Online-Angebote aus Sicht von Usern und Experten. *Medien & Kommunikationswissenschaft*, *53*, 486–506. https://doi.org/10.5771/1615-634x-2005-4-486

Waddell, T. F. (2018). A robot wrote this? How perceived machine authorship affects news credibility. *Digital Journalism*, *6*, 236–255. https://doi.org/10.1080/21670811.2017.1384319

Wissenschaft im Dialog. (2023). *Wissenschaftsbarometer 2023* [Science barometer 2023]. https://wissenschaft-im-dialog.de/projekte/wissenschaftsbarometer/#erhebung-2023

World Health Organization. (2019). *Ten threats to global health in 2019*. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

Yildirim, I., & Paul, L. A. (2024). From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences*, *28*, 404–415. https://doi.org/10.1016/j.tics.2024.02.008

Zhao, X., Strasser, A., Cappella, J. N., Lerman, C., & Fishbein, M. (2011). A measure of perceived argument strength: reliability and validity. *Communication Methods and Measures*, *5*, 48–75. https://doi.org/10.1080/19312458.2010.547822

Zheng, Y., Zhong, B., & Yang, F. (2018). When algorithms meet journalism: the user perception to automated news in a cross-cultural context. *Computers in Human Behavior*, *86*, 266–275. https://doi.org/10.1016/j.chb.2018.04.046

Zhou, Z., Wang, X., Li, X., & Liao, L. (2023). Is ChatGPT an evidence-based doctor? *European Urology*, *84*, 355–356. https://doi.org/10.1016/j.eururo.2023.03.037

About the authors

Selina A. Beckmann is a research associate at the Department of Communication at Johannes Gutenberg University of Mainz, Germany. Her research interests include science and health communication.

selina.beckmann@uni-mainz.de @selinaabeckmann

Elena Link is an assistant professor at the Department of Communication at Johannes Gutenberg University of Mainz, Germany. In her research, she studies health information-seeking and avoidance behaviors, their influencing factors, and their outcomes.

elena.link@uni-mainz.de @elena-link

Marko Bachl is an assistant professor at the Department of Media and Communication Studies at Freie Universität Berlin, Germany. His research interests include digital research methods as well as political and health communication.

marko.bachl@fu-berlin.de @bachl

How to cite

Beckmann, S. A., Link, E. and Bachl, M. (2025). "'ChatGPT, is the influenza vaccination useful?" Comparing perceived argument strength and correctness of pro-vaccination arguments from AI and medical experts'. *JCOM* 24(02), A04. https://doi.org/10.22323/2.24020204.

ISSN 1824-2049. Published by SISSA Medialab. jcom.sissa.it